CRITICAL REVIEW BY SVATI KUSH

Structural Failure Modes in Clinical AI:

A Four-Layer Taxonomy for System-Level Safety

Javier Morales, MD

SSRN Working Paper Draft, May 2026

Companion to: Satti & Ferguson, The Lead Zero Standard (SSRN, 2026)

Abstract

This review offers a critical examination of Morales (2026), 'Structural Failure Modes in Clinical AI: A Four-Layer Taxonomy for System-Level Safety.' The paper is evaluated across its internal logical coherence, empirical grounding, contribution to existing safety and governance literature, and — centrally — its philosophical implications. Drawing on philosophy of mind, epistemology, ethics of technology, and the philosophy of medicine, this review argues that while Morales makes a substantive and well-structured contribution to clinical AI governance, the framework rests upon several unexamined philosophical assumptions regarding knowledge, agency, responsibility, and the nature of clinical judgment. The review identifies these assumptions, traces their implications, and proposes areas where deeper philosophical engagement would strengthen the taxonomy's normative foundations.

Morales argues that even a technically accurate AI can cause harm if the surrounding system — the interface, workflow, and accountability structures — fails.

What the Review Adds — The Philosopher's Lens

The review agrees with Morales but argues he left big philosophical questions unanswered:

Epistemology — AI produces "proxy coherence": language that sounds like knowledge but isn't. Wittgenstein's point applies — clinical knowledge lives inside a practice, not just sentences.

Tacit Knowledge (Polanyi) — Doctors know more than they can say. A clinician overriding AI without explaining why isn't being irrational — she may be using expertise that can't be articulated.

Moral Responsibility — When AI, vendor, interface, institution, and doctor all contribute to a bad outcome, who is actually blameworthy? Morales identifies the causal chain but doesn't answer the moral question.

Autonomous Agency (Frankfurt) — A doctor signing off on an AI plan under time pressure, without full context, may not be exercising genuine consent — which undermines placing responsibility on her.

Philosophy of Medicine — Clinical judgment is irreducibly about this patient, not statistical populations. An AI that's 95% accurate across a population gives no guarantee for the individual case.

Extended Mind (Clark & Chalmers) — If AI becomes deeply embedded in how doctors think, the boundary between "AI suggestion" and "doctor's decision" may blur ontologically — making Morales's admissibility gates harder to enforce than they appear.

The Verdict

Morales's framework is practically excellent — a useful scaffold for hospitals, regulators, and procurement teams. But philosophically it's normatively thin: it tells you where to put governance gates, not what values those gates should protect or how to distribute moral blame when they fail.

1. Introduction

The deployment of artificial intelligence in clinical medicine has generated an urgent and still-unresolved set of governance problems. Morales (2026) enters this debate with a framework that is at once analytically careful and practically oriented. The paper's central claim — that safety is a property not of the model but of the coupled human-interface-workflow-accountability system — is both theoretically defensible and empirically motivated. It aligns with decades of sociotechnical systems research and extends that tradition into the specific pathology of clinical AI.

Yet a paper that undertakes to locate the site of safety, to define what counts as failure, and to prescribe governance accordingly is, whether explicitly or not, engaged in philosophical work. It makes claims about knowledge and its limits (epistemology), about what clinical judgment is and who possesses it (philosophy of mind and medicine), about how responsibility should be assigned across human and non-human actors (ethics and moral philosophy), and about the relationship between rules, systems, and action (philosophy of governance and social ontology).

This review takes those philosophical commitments seriously. Section 2 summarises the paper's argument and structural contributions. Section 3 evaluates the framework from the perspective of epistemology and philosophy of science. Section 4 examines the paper's implicit theory of agency and moral responsibility. Section 5 addresses the philosophy of medicine and the nature of clinical judgment. Section 6 considers the paper through the lens of systems ontology and sociotechnical theory. Section 7 identifies limitations. Section 8 offers a synthesising assessment.

2. Summary of the Paper's Argument

2.1 Core Thesis

Morales advances a systems-level critique of model-centric AI safety discourse. The paper argues that even a well-aligned, accurate, and explainable model can participate in unsafe clinical outcomes when the surrounding system — interface, workflow, accountability structures — fails to preserve the conditions under which model output may legitimately enter clinical action. The animating insight is that fluency is not validity, and that authority is conferred by systems, not by accuracy alone.

2.2 The Four-Layer Taxonomy

The paper organises its analysis into four analytically distinct but operationally coupled layers:

Layer	Core Question	Primary Failure Mode	Governance Target
Representation	Does model output track clinical reality?	Proxy coherence: fluent output detached from physiology	Ground uncertainty; mark source and evidential limits
Interface	How does interaction reshape meaning and authority?	Authority inflation through framing, tone, or role assumption	Preserve uncertainty; distinguish suggestion from order
Integration	What happens when output enters workflow?	Automation bias, anchoring, silent propagation	Admissibility gates; workflow-aware human review
Accountability	Who controls, signs, audits, and bears responsibility?	Influence without corresponding control or traceability	Preserve clinician authority; define vendor responsibility

2.3 Extended Failure Modes

Beyond the four-layer structure, Morales identifies a set of interpretive and temporal failure modes of particular salience in high-complexity human-AI interaction. These include temporal aliasing (misreading expert inferential speed as instability), longitudinal calibration blindness (evaluating reliability from a single snapshot rather than a trajectory), snapshot pathology inference (diagnosing instability from isolated utterances), metaphor literalization failure (treating heuristic language as literal belief), neurodivergent calibration mismatch (applying inappropriate norms to non-standard communication styles), and execution boundary collapse (the unmarked passage from AI-generated information to clinical action).

2.4 The Central Governance Shift: From Alignment to Admissibility

The paper's most distinctive contribution is the distinction between alignment and admissibility. Alignment asks whether model behavior is acceptable. Admissibility asks whether a given output is permitted to enter clinical action under defined conditions. This reframes the governance problem: rather than asking 'Is the model safe?', institutions must ask 'Under what conditions may this output act?'

3. Epistemological Analysis

3.1 Proxy Coherence and the Limits of Linguistic Knowledge

Morales's concept of proxy coherence — the appearance of reasoning produced by linguistic structure without attachment to clinical reality — is epistemologically rich and philosophically consequential. It is, in essence, a claim about the difference between syntactic and semantic adequacy: a model may produce outputs that are linguistically well-formed and contextually coherent without possessing the kind of world-tracking relationship that constitutes genuine knowledge.

This resonates strongly with Wittgenstein's later philosophy of language, particularly the notion that meaning is use within a form of life (Wittgenstein, 1953). A clinical statement does not derive its meaning from its internal structure alone; it acquires force within a practice — the practice of medicine — constituted by training, embodied judgment, institutional norms, and patient-specific context. A language model reproduces the surface of that practice without participating in its constitutive form of life. The epistemological implication, which Morales gestures toward but does not fully develop, is that clinical knowledge is not propositional in the sense that AI outputs are propositional. It involves a kind of knowing-how — what Ryle (1949) calls procedural knowledge — that cannot be fully captured or transmitted through linguistic artifacts.

This is relevant to governance because it suggests that the 'verification' Morales prescribes at admissibility gates may be harder to specify than the framework implies. A clinician who reviews an AI output must not merely confirm its factual accuracy; she must integrate it into a form of knowing that language models cannot replicate. The admissibility framework would benefit from a more explicit account of what clinician review actually consists in, epistemologically speaking.

3.2 Longitudinal Calibration as Epistemic Virtue

The concept of longitudinal calibration — evaluating reliability across time through error correction, responsiveness to evidence, and coherence of trajectory — can be mapped productively onto virtue epistemology, particularly the work of Zagzebski (1996) and Sosa (2007). On these accounts, epistemic reliability is not a property of individual beliefs or outputs but of cognitive processes that characteristically produce truth across contexts. What Morales calls calibration is close to what virtue epistemologists call intellectual virtue: a stable disposition to reason well, correct errors, and track evidence.

The virtue-epistemological framing has a practical implication that Morales does not make explicit: if clinical AI systems are to be evaluated longitudinally, the evaluative process itself must possess certain virtues — patience, contextual sensitivity, freedom from confirmation bias, and attentiveness to correction dynamics. A governance process that lacks these virtues will miscalibrate even a well-calibrated system. This points to a second-order epistemological problem: the governance of AI systems requires epistemic virtue in the governors, not merely technical criteria in the assessed.

3.3 The Problem of Tacit Knowledge

Both the representation layer failure (proxy coherence) and the interface layer failure (expertise invisibility) implicitly depend on a theory of tacit knowledge. Morales notes that 'high-compression expert reasoning is dismissed because intermediate steps are not externally visible.' This is precisely what Polanyi (1966) describes as tacit knowing: expert competence that exceeds its own articulation. We know more than we can tell.

The governance implication is significant. If clinical expertise is partly constituted by tacit knowledge that resists explicit representation, then any admissibility framework that relies solely on articulable criteria will systematically undervalue expert override of AI outputs. Conversely, a clinician's failure to articulate her reasons for rejecting an AI recommendation should not be interpreted as cognitive failure. Governance structures should create protected space for non-articulable clinical judgment, not merely incentivise compliance with AI-generated pathways.

4. Agency, Responsibility, and the Ethics of Distributed Systems

4.1 The Problem of Many Hands

Morales's accountability layer identifies what philosophers of technology call the problem of many hands (van den Hoven, 2012; Thompson, 1980): when harm results from the distributed contributions of many actors — model developer, vendor, interface designer, institution, and clinician — moral responsibility becomes diffuse to the point of effective disappearance. This is not unique to AI; it is a structural feature of complex sociotechnical systems. But clinical AI intensifies it because the downstream stakes include patient harm, professional liability, and public trust in medicine.

Morales's governance response — traceability, audit trails, explicit authorization boundaries — is sensible but philosophically incomplete. Traceability is a necessary condition for accountability, but it is not sufficient. Knowing who did what does not resolve the question of how responsibility should be distributed. Should a clinician bear full responsibility for an AI-influenced decision she made under time pressure with inadequate information about the model's limitations? Should a vendor bear responsibility for foreseeable automation bias that they chose not to mitigate? The paper leaves these normative questions structurally open.

A fuller philosophical treatment would draw on the distinction between causal responsibility (who contributed to the outcome) and moral responsibility (who is blameworthy and to what degree), as developed by Fischer and Ravizza (1998) and Strawson (1962). Morales's framework identifies causal chains effectively. It does not articulate a principle for converting causal contribution into moral blame or institutional duty — a gap that matters considerably for legal and professional standards.

4.2 Can AI Systems Bear Responsibility?

Morales is careful to frame AI outputs as information artifacts that must be authorised by human actors before acquiring clinical consequence. This implicitly endorses what Floridi and Cowls (2019) call a 'patient-not-agent' view of AI: systems can be objects of moral consideration but not bearers of moral responsibility. The clinician signs the order; the model does not.

This is a defensible position, but it becomes strained in agentic and semi-agentic configurations — which Morales explicitly acknowledges as within the paper's scope. When an AI system not only generates a recommendation but initiates a task plan, delegates steps across tools, and executes actions, the question of whether it functions as a moral agent becomes practically urgent. Luciano Floridi's work on moral agency at the level of artificial agents (Floridi & Chiriatti, 2020) suggests that sufficiently autonomous systems may require a form of functional moral status, even absent sentience. Morales's

admissibility framework may need to be extended to account for systems that cross the threshold from information to action in ways that precede human review.

4.3 Responsibility and the Conditions of Meaningful Consent

The paper's insistence that 'accountability cannot be deferred to the clinician while control is distributed elsewhere' raises a further philosophical question: what conditions must be satisfied for a clinician's decision to count as genuinely autonomous — and therefore as a legitimate ground for responsibility? Following Frankfurt (1971) and Bratman (1987), autonomous agency requires not merely choosing among options but doing so on the basis of second-order volitions: reflective endorsement of one's first-order desires and reasons.

A clinician who signs an AI-generated plan under severe time pressure, without adequate information about the model's evidential basis, and within an institutional culture that rewards AI adoption, may not be exercising autonomous agency in the morally relevant sense. Morales's governance framework should include conditions specifying when a clinician's review constitutes genuine autonomous authorization, as opposed to performative compliance with institutional expectation.

---

5. Philosophy of Medicine and the Nature of Clinical Judgment

5.1 What Is Clinical Knowledge?

The paper's distinction between probabilistic model output and clinical reality implicitly invokes a longstanding debate in the philosophy of medicine between nomothetic and idiographic conceptions of clinical knowledge (Feinstein, 1967; Wulff, Pedersen, & Rosenberg, 1986). Nomothetic knowledge is knowledge of general laws and statistical regularities — the domain in which large language models are trained. Idiographic knowledge is knowledge of the particular case: this patient, at this moment, in this context.

Morales's proxy coherence failure is precisely the confusion of the nomothetic for the idiographic: a model that accurately represents population-level patterns may systematically misrepresent any individual case that deviates from the population mean. This is not merely a technical limitation; it is an epistemological one. The idiographic knowledge required for clinical judgment is not derivable from nomothetic training data, however extensive.

The governance implication is that admissibility criteria should not treat model accuracy at the population level as a sufficient condition for admissibility at the patient level. A model that is 95% accurate across a population provides no specific reassurance for the patient who falls in the 5%. Morales's framework needs a more explicit account of how admissibility decisions should handle this idiographic gap.

5.2 Clinical Judgment as a Moral Practice

The philosophy of medicine has increasingly emphasized that clinical judgment is not merely epistemic but moral: it involves not only 'what is the case?' but 'what should be done?' and 'who has the authority and responsibility to decide?' (Pellegrino & Thomasma, 1981; Montgomery, 2006). Morales's accountability layer captures part of this by insisting on preserved clinician authority. But the paper does not explicitly address the fiduciary nature of clinical responsibility — the physician's duty to the patient that is irreducible to institutional compliance.

This is philosophically significant because fiduciary obligations cannot be transferred or delegated in the way that information tasks can. A physician who delegates the generation of a care plan

to an AI system does not thereby delegate her moral obligation to the patient. The question of whether AI-mediated decisions are compatible with fiduciary duty is not merely regulatory but normative: it concerns what medicine is, not only how it is regulated.

---

6. Systems Ontology and Sociotechnical Theory

6.1 The System as the Unit of Safety

Morales's foundational claim — that safety is a property of the system — reflects a systems-theoretic ontology in which the relevant unit of analysis is not the component but the configuration of relationships among components. This is consonant with Leveson's STAMP (Systems-Theoretic Accident Model and Processes) and with Reason's Swiss Cheese Model, both of which Morales cites. Philosophically, it resonates with emergence theory: system-level properties (safety, failure) cannot be reduced to or predicted from component-level properties.

What this implies for governance is that regulatory approaches that focus exclusively on model properties — accuracy, explainability, bias — are systematically inadequate because they treat the model as the relevant unit. Morales's taxonomy is valuable precisely because it shifts the unit to the system. However, the paper stops short of addressing a deeper ontological question: where are the boundaries of the system? The taxonomy identifies four layers, but in practice, clinical AI systems are embedded within larger institutional, professional, legal, and cultural systems whose influence on failure propagation is real but not addressed.

6.2 Distributed Cognition and the Extended Mind

The integration layer's account of human-AI coupling implicitly draws on distributed cognition theory (Hutchins, 1995) — which Morales cites — and is consistent with the extended mind hypothesis of Clark and Chalmers (1998). On this view, cognitive processes are not confined to the individual skull but extend into the environment, including into tools, interfaces, and collaborative systems. If clinical reasoning is partly constituted by AI tools, then the question of where the clinician's cognition ends and the AI's output begins becomes genuinely unclear.

This has a governance implication that the paper does not fully develop: if AI tools become genuinely constitutive of clinical cognition — as opposed to merely informing it — then the admissibility framework's distinction between information and action may be harder to maintain than the paper suggests. The extended mind thesis implies that the boundary Morales wishes to enforce (between AI suggestion and clinical decision) may be ontologically unstable, not merely difficult to implement.

---

7. Critical Limitations

7.1 Absence of Normative Foundations

The paper's most significant philosophical limitation is its failure to ground its governance prescriptions in an explicit normative framework. Morales identifies failure modes and proposes controls, but does not articulate the underlying values that the controls are designed to protect. Patient autonomy? Clinical efficacy? Professional integrity? Institutional accountability? These values can and do conflict, and a taxonomy that does not specify their priority ordering is vulnerable to being applied in ways that optimise for some values at the expense of others.

7.2 The Admissibility Concept Requires Thicker Specification

The distinction between alignment and admissibility is conceptually valuable. However, admissibility as defined in the paper — 'the criteria governing whether an AI-generated output may enter clinical workflow or action' — is a formal rather than a substantive concept. It tells us that criteria are needed; it does not tell us what the criteria should be, how they should be determined, or who has the authority to set them. Filling this gap requires normative theory that the paper does not supply.

7.3 Empirical Validation Absent

The paper is explicitly a conceptual contribution. It makes no empirical claims and presents no case studies, incident analyses, or simulation data. This is acknowledged in the limitations section. However, it means that the failure modes, as named and defined, remain taxonomic hypotheses rather than empirically established categories. The degree to which the proposed taxonomy maps onto real clinical AI incidents — as opposed to conceptually possible ones — is unestablished.

7.4 The Clinician as Residual Bearer of Risk

Despite the paper's insistence that accountability cannot be deferred to the clinician while control is distributed elsewhere, the operational controls proposed tend to culminate in clinician review and authorization. In practice, this risks reproducing the very asymmetry the paper diagnoses: the clinician becomes the last point of accountability in a system that has already constrained her choices through framing, workflow embedding, and documentation defaults. A more radical governance response — in which vendor accountability is primary and clinician authorization is final rather than residual — would address this asymmetry more directly.

---

8. Synthesis and Overall Assessment

Morales (2026) makes a substantive and timely contribution to clinical AI governance. Its four-layer taxonomy provides a practically useful framework for identifying where alignment breaks across the system, and its extended failure modes add analytical granularity where existing literature has been imprecise. The central distinction between alignment and admissibility is the paper's strongest conceptual contribution and deserves wide adoption in clinical AI evaluation, procurement, and regulatory practice.

Philosophically, the paper's core claim — that safety is systemic, not model-local — is well-founded and resonates with epistemological, ethical, and ontological frameworks that the paper largely leaves implicit. The present review has attempted to surface those frameworks and trace their implications. The most significant philosophical gaps are:

- The absence of an explicit normative foundation specifying the values that the taxonomy is designed to protect and their priority ordering.
- The treatment of clinical judgment as an evaluative endpoint without adequate philosophical specification of what that judgment consists in, epistemologically and morally.
- An underdeveloped account of moral responsibility distribution, which leaves the framework procedurally complete but normatively thin.
- Insufficient engagement with the ontological instability of the information-action boundary in agentic systems — a boundary on which the admissibility framework depends.

None of these limitations undermine the paper's practical utility. They do suggest that the taxonomy, as presented, is better understood as a governance scaffold than as a complete theory of clinical AI safety. A fuller theory would need to engage philosophy of mind, virtue epistemology, the

ethics of distributed agency, and the philosophy of medicine more explicitly — not to make the framework more academic, but to make it more robust when instantiated in the contested institutional environments where clinical AI is actually deployed.

> *Safety is not a property of the model. It is a property of the system, and whether its structure holds under pressure. — Morales (2026, p. 7)*

This conclusion is philosophically sound. The further question, which this review has attempted to open, is: what structural properties must a system possess not only to prevent unsafe outputs but to do so in ways that are epistemically defensible, ethically legitimate, and institutionally sustainable? That question is beyond the scope of any single paper. But Morales has made a careful and serious contribution toward answering it.

References

Bratman, M. E. (1987). Intention, Plans, and Practical Reason. Harvard University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7-19.

Feinstein, A. R. (1967). Clinical Judgment. Williams & Wilkins.

Fischer, J. M., & Ravizza, M. (1998). Responsibility and Control: A Theory of Moral Responsibility. Cambridge University Press.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. Journal of Philosophy, 68(1), 5-20.

Hutchins, E. (1995). Cognition in the Wild. MIT Press.

Leveson, N. (2004). A new accident model for engineering safer systems. Safety Science, 42(4), 237-270.

Montgomery, K. (2006). How Doctors Think: Clinical Judgment and the Practice of Medicine. Oxford University Press.

Morales, J. (2026). Structural failure modes in clinical AI: A four-layer taxonomy for system-level safety. SSRN Working Paper.

Pellegrino, E. D., & Thomasma, D. C. (1981). A Philosophical Basis of Medical Practice. Oxford University Press.

Polanyi, M. (1966). The Tacit Dimension. Doubleday.

Reason, J. (2000). Human error: Models and management. BMJ, 320(7237), 768-770.

Ryle, G. (1949). The Concept of Mind. Hutchinson.

Satti, E., & Ferguson, J. (2026). The Lead Zero Standard: A governance protocol for clinician-as-originator architecture in AI-assisted medicine. SSRN Working Paper.

Sosa, E. (2007). A Virtue Epistemology: Apt Belief and Reflective Knowledge. Oxford University Press.

Strawson, P. F. (1962). Freedom and resentment. Proceedings of the British Academy, 48, 1-25.

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. American Political Science Review, 74(4), 905-916.

van den Hoven, J. (2012). The use of normative theories in computer ethics. In L. Floridi (Ed.), The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press.

Wittgenstein, L. (1953). Philosophical Investigations. Blackwell.

Wulff, H. R., Pedersen, S. A., & Rosenberg, R. (1986). Philosophy of Medicine: An Introduction. Blackwell.

Zagzebski, L. T. (1996). Virtues of the Mind. Cambridge University Press.

Dear Dr. Morales,

I write to express my sincere gratitude for the opportunity to review your working paper, ‘Structural Failure Modes in Clinical AI: A Four-Layer Taxonomy for System-Level Safety.’

Your framework is a substantive and timely contribution to clinical AI governance. The distinction between alignment and admissibility, in particular, clarifies a conceptual gap that much of the existing literature has left unaddressed. Engaging with your work prompted a deeper examination of the epistemological and ethical dimensions underlying systems-level safety — an exercise I found genuinely enriching.

I hope the review proves useful as you develop the paper further. I would welcome any opportunity to discuss the philosophical threads it raises.

With respect and appreciation

Svati kush

About the Writer

Svati Kush is a New Delhi-based poet, essayist, and literary critic whose writing refuses to separate the aesthetic from the political, the ancient from the contemporary, or the personal from the civilizational.

Her essays have been acknowledged by Dr. Shashi Tharoor on Twitter. She is the author of Universal Memories: A Book of Poetry of Life. Her first book-length critical study — Seven Silence: A Poetic Journey Through R.K. Singh's Universe — was written at the personal invitation of Prof. R.K. Singh, internationally translated poet and Pushcart Prize nominee, who noted that the work has the potential to serve future scholars of Indian English poetry.

Literature is not decoration. It is diagnosis.

And I intend to keep writing prescriptions.